

DOCUMENT RESUME

ED 126 121

TE 005 376

AUTHOR Polley, Freda S.; And Others
TITLE Models for the Delivery of School District Evaluation: Service or Accountability?
PUB DATE [Apr 76]
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)
DESCRIPTORS MF-\$0.83 HC-\$1.67 Plus Postage.
Decision Making; Educational Accountability; *Evaluation Methods; *Evaluation Needs; *Models; Organization; *Program Evaluation; *School Districts

ABSTRACT

The arguments for evaluation as a service versus evaluation as an accountability function are debated. Proponents of the evaluation as service approach often argue that information requested for self-renewal is the information most likely to be used while those who favor accountability as the appropriate role for evaluation claim that people will seek out evaluation of themselves only if it is not dangerous to them personally. These points and others are argued point and counterpoint; a resolution between the two views is sought. (Author)

* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *

ED126121

Models for the Delivery of School District Evaluation:
Service or Accountability?

Freda M. Bailey,
Jack L. Davidson, Ann M. Lee, and Paula Matuszek

Paper Presented at the Annual Meeting of the American Educational Research Association

San Francisco

April, 1976

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

"THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY."

TM005 376

Models for the Delivery of School District Evaluation:

Service or Accountability?

Board of Trustees

Carole Keeton McClellan, President

Gustavo L. Garcia, Vice President

Reverend Marvin C. Griffin, Secretary

Will Davis

DeCourcy Kelley

Jerry Nugent

M. K. Hage, Jr.

Superintendent of Schools

Dr. Jack L. Davidson

Director, Office of Research and Evaluation

Dr. Freda M. Holley

Models for the Delivery of School District Evaluation:
Service or Accountability?

Fuzzy thinking underlies many of the organizational schemas for evaluation in public schools today—fuzzy thinking about the purposes of evaluation and the underlying philosophies that govern the organizational structure of evaluation units to achieve those purposes. The way a school district organizes to do evaluation should directly reflect both its purpose in evaluating its program and its philosophies of administration and educational change. Evaluators themselves should be keenly attuned to these local ideas, either directing their work to the existing purposes and philosophies or working to change them.

The Austin ISD Office of Research and Evaluation was created three years ago in a district where the Board and public were clamoring for accountability. In reexamining our actions over the past three years as we have struggled to institutionalize this new unit, we believe we have been guilty of the fuzzy thinking alluded to above. We tried to serve up a "service" unit when the dish our paying clients were ordering was an "accountability" one. In fact, we found that we were unclear three years ago about who our real clients were and hence were misaddressing much of our work and reporting. Moreover, we think we were deluding ourselves with some romantic notions about human behavior change that the data simply will not support.

In the last three years, we have laid out two dichotomous models of "evaluation for accountability" and "evaluation as a service function." By examining these two models, we have come to better understand our organizational role. This self-examination has had concomitant effects on the focus of our work, the establishing of priorities, the external and internal organization of our unit, and particularly on our reporting formats and styles. Most of all it has helped our evaluation unit have an impact on educational practice in our district.

This paper first presents some concepts that underpin the two models to be described. Then it lays out the two models. These models are described first as two extremes, probably neither of which exist in pure form in any school district. A third model, a compromise of these two extremes, which can result from an analysis of these first two models is portrayed.

CONCEPTUAL FRAMEWORK

The Client for Evaluation

Client is a crucial term. Clients may be one of two types: client-purchaser or client-recipient. The distinguishing feature between these two types of clients is that the client-purchaser has direct power to purchase while the client-recipient may receive goods or services only if he influences those who have purchase-power to procure the services for him. Using this definition, the client-purchaser will be the administration (superintendent, board, and public) while the client-recipient will be the school personnel who might use the evaluation service for their purposes.

Evaluation Purpose

We define the purpose of evaluation as the use or potential use to which evaluation may be addressed. Our office identifies the ultimate purpose of evaluation as the improvement of student learning outcomes.

All clients, whether paying or using, will agree theoretically with this definition. Interpediate purposes, however, may not be so clear. While it is true that the public, school board, and superintendent may well have political evaluation purposes in mind, the school district staff may be even more "politically" motivated, because they have at stake the very personal purpose of professional advancement or job retention. Thus, evaluators are always faced with a political reality in which negative evaluation information is rarely well received. Glass (1975) has made us aware of the paradoxical climate in which evaluation information is generated and used.

Despite the problems associated with defining the purpose of evaluation, the wise evaluator should understand the environment in which he operates and the potential purposes for evaluation with both kinds of clients identified.

Philosophy and Evaluation

Throughout the history of man, two conflicting views of human nature have operated. In one view, man is viewed as innately good. As such, he will choose to behave rationally and kindly toward his fellow man, unless society (viewed as an evil corrupter) influences him to behave wrongfully. In the alternative view, man is seen as born with "original sin." Here, man is viewed as requiring exorcism of evil through baptism, training, or other continued vigilance. More modern philosophers may phrase this as innate self-centeredness. In education the basically good vision of man is expressed in the philosophies of Rousseau and such contemporary romanticists as Holt, Kohl, and Illich. The evil view of man is exemplified in the "world view" of Newton (Mink) and more recently the "fundamental school" and "basic skills" citizen groups.

although the relationship may appear oblique, these philosophical dichotomies can be seen underlying the two models of evaluation as accountability or service.

TWO MODELS OF EVALUATION

1. Accountability Model

The clients for evaluation here are clearly client-purchasers—the top level administrators, board, and thus indirectly the broader public. Their need of an evaluation unit is for a brief, reliable report on whether the program worked or didn't work in terms of products or student outcomes in order to make decisions. They will be particularly interested in cost information. Not only will they be interested in total program costs, but if there were achievement gains they will wish to know at what relative costs they were gained. They will need enough written backup data to be sure of the evaluation unit's competence, but they will not wish to hear this data in detail. If there is any process information of direct importance to outcomes, such as a major failure in the implementation of the program, they may be interested in knowing this. But they will not want to hear about the fine points of the process evaluation.

The philosophy underlying this model is that people act in response to directions from an authority and usually must be externally motivated to change their performance. Contrary to much of the recent human resources and systems analysis theorizing, this model postulates that in the complex world of human behavior, simple feedback on the state of the system will not be sufficient to change behavior. The immediate human costs for change are frequently so great that external intervention and orders from above are necessary.

Organizationally, the essentials that must be observed if the program accountability model is elected is that the evaluation should be an independent organizational unit reporting at a very high level of the hierarchy. This is necessary, because credibility is the most crucial commodity an evaluator has in this environment. The evaluator's clients will not have time to read his massive technical reports nor understand his fine statistics, but they must somehow be assured of his competence and integrity. Moreover, they must have confidence that his work will not be filtered of negative information. Thus, in the diagram of this model in Figure 1 the evaluation unit is hierarchically above the program unit. This position of the unit will also be crucial, because the total system is likely to become the natural enemy (overt or covert) of the unit.

INSERT FIGURE 1

If this accountability model of evaluation were strictly implemented, it would be the least expensive form of evaluation in monetary terms. Since communication and coordination with program or school staff and process evaluation tend to be the most time consuming and hence most expensive elements of evaluation, the reduction in importance of the downward flow of information in this model can reduce evaluation costs.

In terms of staff support, however, the approach strictly implemented will be the most "expensive." Staff resentment over the evaluator (as auditor) will gradually lead to distrust. This distrust in turn will interfere with data access and/or data reliability.

2. Service Model

The client in this case is the client-receiver. (usually the program staff), and evaluation assumes a service role to that staff. In the

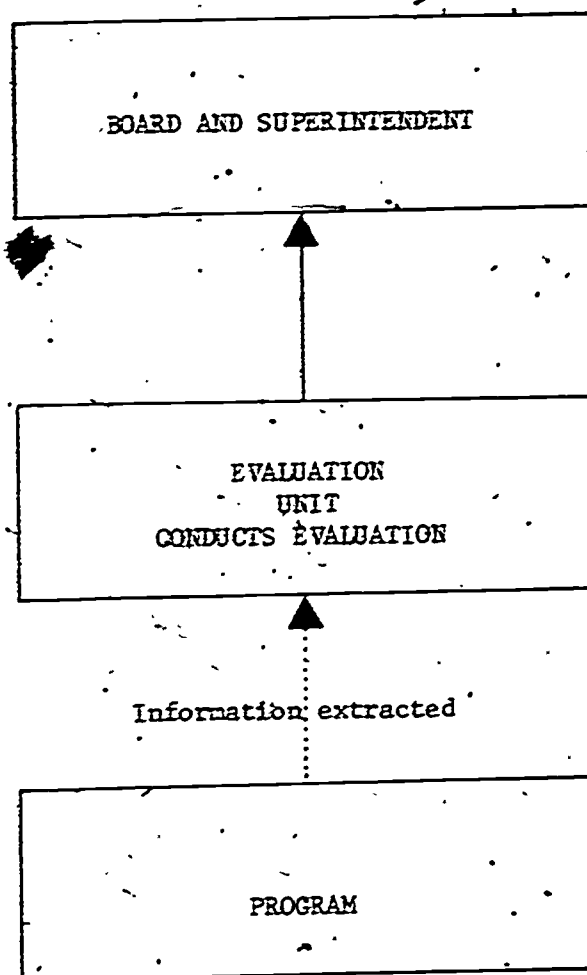
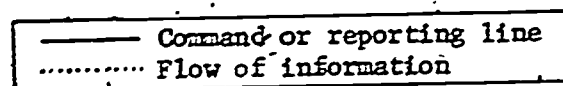


Figure 1: ACCOUNTABILITY MODEL OF EVALUATION



service model's purest form, the evaluation staff might provide to the staff only technical assistance such as statistical analyses or data processing. The evaluation staff might even report to the program staff administratively. This organizational arrangement is illustrated in Figure 2. The two-way directionality of information flow in this model as opposed to the previous diagram is the most important feature.

INSERT FIGURE 2

This model has its philosophical roots in the old rousseauian view of human nature: it is believed that man will opt for "good" under his own motivation--in this case, educational improvement. The underlying postulate is that people are always motivated to perform at their maximum level; evaluation serves only to provide accurate information about the state of the program or system. It is implied that no external motivation or threat is needed to provide change.

The evaluation staff following this model would be quickly attuned to in-course program changes and could pick up on process differences as they occur. Ideally, evaluation services would be requested at appropriate times, and the data would be immediately used by the program staff.

Several weaknesses are immediately obvious in this model. Most importantly, negative information is likely to be filtered out before reaching major decision-makers. This may lead to the retention of bad or weak programs. By the same token, the evaluation design may be at the mercy of the program staff. The clients may not request or approve appropriate evaluation resources, which will then lead to inaccurate or invalid data or its interpretation. The program staff is also unlikely to place a high priority on the long-term pay-offs of evaluation compared to immediate budget needs of the program. Hence, the allocation

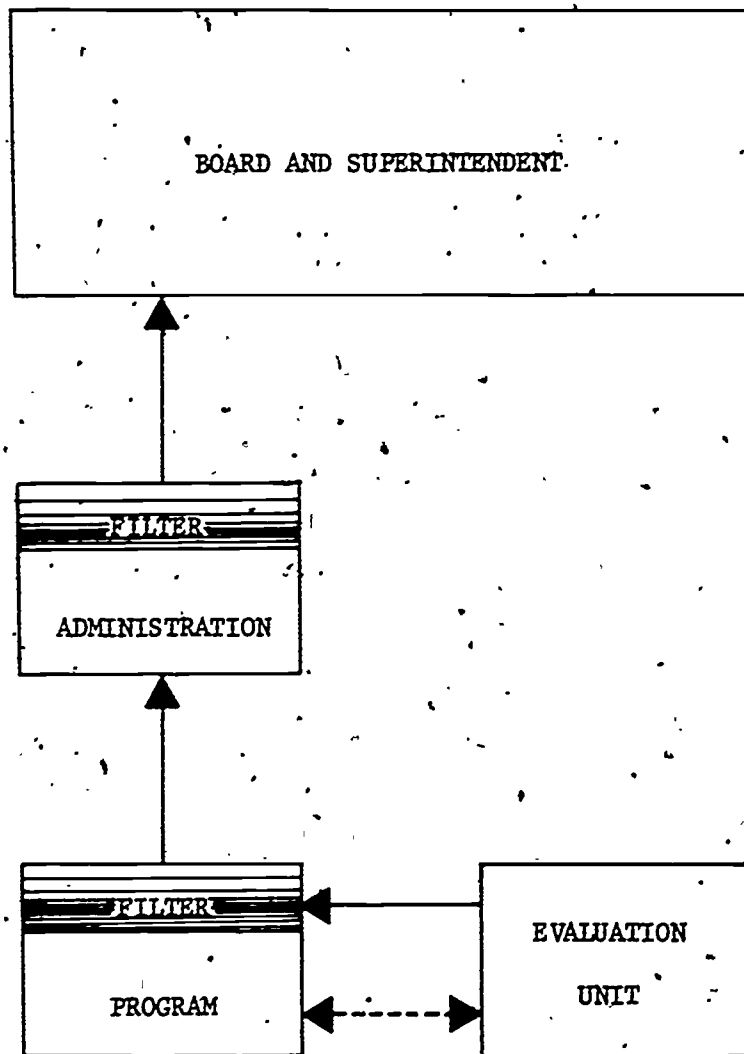
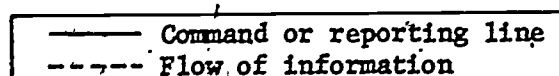


Figure 2: SERVICE MODEL OF EVALUATION



of sufficient resources to evaluation to provide adequate process and outcome evaluation will be unlikely. Inadequate evaluations can then lead to evaluation receiving even lower resource priority, and the downward spiral continues. Perhaps the most serious problem for this model for public school evaluators is that the information client here is not the money-paying client who must eventually support the unit with tax dollars.

THESE TWO MODELS AND REALITY (A THIRD MODEL)

Extensive literature on evaluation theory has appeared addressing such questions as the function and purposes of evaluation (Stufflebeam, et al., 1970, Provus, 1971, Stake, 1967) and the methodology of evaluation (Popham, 1974, Borich, 1974, Anderson et. al., 1975). Many of these earlier works have tangentially dealt with the internal organization of evaluation units, but none have really made explicit the alternate organizational ways in which evaluation services might be delivered within a local school district.

We believe these models of evaluation as service and as accountability may have more practical implications for the organizational realities faced by public school evaluation staffs than other extant theoretical work. Regardless of how well or how thoroughly evaluations are conducted, the translation of those evaluations into program actions is often more dependent upon the organizational role which evaluation plays than upon the study itself. We have seen numerous evaluations dropped into the great chasm of public school bureaucracy never to be heard from again because the evaluators lacked the organizational voice to have them heard.

The resolution of the competing concepts of human nature and behavior change which are implicit in the dichotomous models described above are not likely to be totally resolved anywhere and certainly not in the political

context in which public school evaluation occurs today. Thus, one is unlikely to see any school district evaluation unit organization that matches the models described. Indeed, most units will try to achieve a working blend of the two models just as we have over our three years of operation.

In the 1973-74 school year our unit began under an ESEA Title III grant to provide a model evaluation capability in our district. Again with fuzzy thinking predominant in our first work, we hoped to achieve accountability through providing information as a service to programs on the achievement of their objectives. We ranked low in the school district hierarchy that first year, reporting to a director who reported to an assistant superintendent who reported to a superintendent. Fortunately, we were established organizationally independent from the programs we evaluated, in direct opposition to some local administrative thinking. At the end of that year we communicated to the program staffs on how well they had done with their objectives. We did manage to get a school board review of those reports, but failed to secure any kind of staff or administrative commitment to program change. Not surprisingly, little program change occurred. The next year's funding from the district for evaluation was about the same--no substantial change.

By whatever yardsticks we could use to measure the effect of our evaluations on district programs that year, we had to rate our unit as a failure. At least, however, we were beginning to recognize what we should be doing.

The second year, we began to take a look at our own work and to analyze our operation and our reports. Thanks also to some fortunate organizational changes that occurred, we reported that second year to a

newly established deputy superintendent charged with overall responsibility for instruction and development. We decided that the pure "service" model, ~~per se~~, was not a viable model and opted for a move toward the accountability mode. In reporting, this meant we emphasized not objectives but decision questions in our reports. An example is indicated in Figure 3.

INSERT FIGURE 3

We forgot our research "if and maybe" conclusions and opted for firm, best-indicated-course recommendations. Moreover, we began to talk face to face to board members and top administrators and to recognize that reports were less important than our personal availability when decisions were about to occur.

In this our third year, we have directly reported to the superintendent and recognize once again the greater impact of evaluation findings on action when input is given directly to the top administrative levels. We believe that our move toward an accountability model has demonstrated greater payoff for changing educational programs. Our current model of operation might be described as in Figure 4.

INSERT FIGURE 4

Nonetheless, we cannot say that a pure accountability model is viable either. "Service," after all, buys access to information, and any evaluation unit that hopes to continue functioning has to yield enough service to keep its data channels open.

Thus, we think we have come up with a blend of the "Accountability-Service Model" in which we acknowledge the true role of the "service" we can render. We do not believe that feedback of evaluation data alone will bring about change in people or in education.

Decision Questions to be Addressed by the IGE Evaluation

A. System-Level Decision Questions

Answers to these questions will assist the Board of Trustees and the Superintendent in making decisions relative to the continuance of the program.

1. Should the Individually Guided Education (IGE) Program be continued at its present level, expanded, or discontinued?
2. Should the IGE Program be implemented only in schools with student groups having certain identifiable characteristics?
3. Are additional resource inputs advisable if the decisions on the first and second questions are positive?
4. Are there any particular characteristics of IGE or individualized instruction whose implementation should be encouraged in AISD elementary schools?

B. Program-Level Decision Questions

Answers to the following questions will assist those charged with implementing the program in their decision making.

1. Should additional training be provided?
2. If additional training is required, should it be of a particular type?

C. School and Classroom Level Questions

Answers to the questions below will assist those charged with making decisions at the school and classroom level, e.g., principals and teachers.

1. Should adaptations or changes be made in the model processes of IGE as implementation proceeds?
2. Is additional resource help needed for the IGE implementation?
3. Do particular staff members or units need additional training or assistance in the implementation of the IGE program?
4. Are there certain types of students who may need to be given particular attention in the IGE classroom?

Figure 3: AN EXAMPLE OF A DECISION QUESTION TABLE

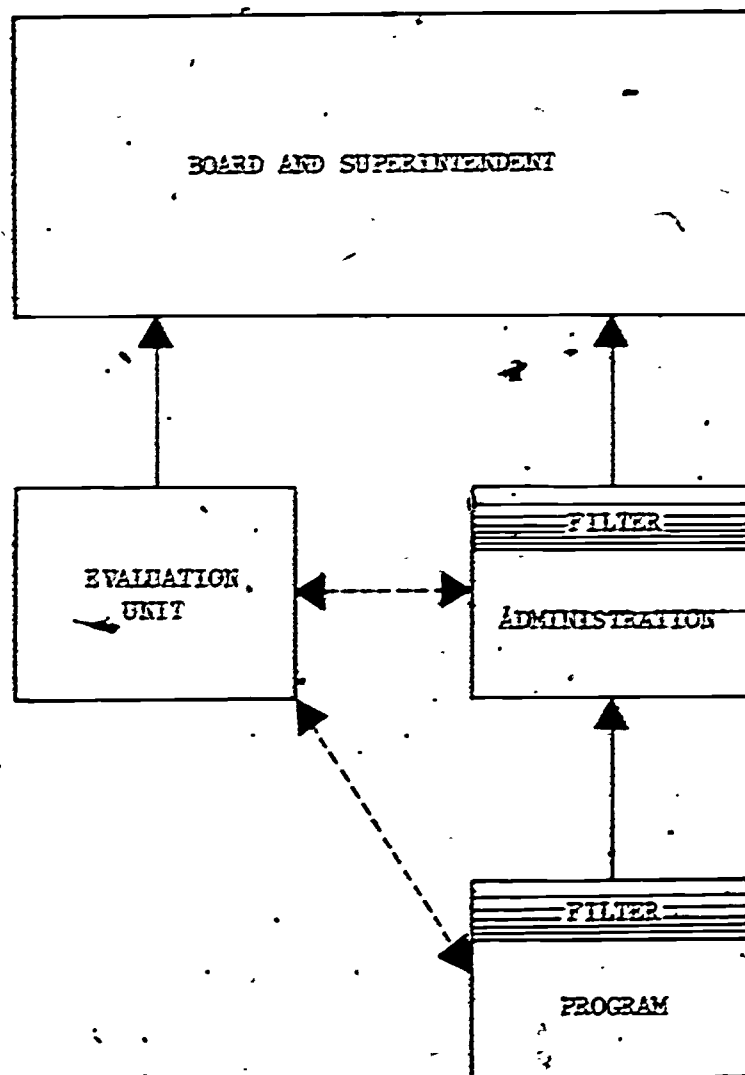
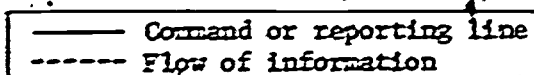


Figure 4: ACCOUNTABILITY-SERVICE-COMPROMISE MODEL
OF EVALUATION



References

Anderson, Scarvia B., et al. Encyclopedia of Educational Evaluation.

San Francisco: Jossey Bass, Inc., 1975.

Borich, Gary D., ed. Evaluating Educational Programs and Products.

Englewood Cliffs, NJ: Educational Technology Publications, 1974.

Glass, Gene V. A Paradox About Excellence of Schools and the People
in Them. Educational Researcher. March, 1975, 4, 9-13.

Mink, Oscar. The Psychology of Planning. Mimeo paper, N.D., p. 5.

Popham, W. James, editor. Evaluation in Education. Berkeley, CA:

McCutcheon Publishing Company, 1974.

Provus, Malcolm. Discrepancy Evaluation Model. Pittsburgh, PA:

Pittsburgh Public Schools, 1969.

Stake, R.E. The Countenance of Educational Evaluation. Teachers College
Record, 1967, 68, 523-40.

Stufflebeam, David L., et al. Educational Evaluation and Decision
Making. Itasca, IL: F. G. Peacock Publishers, Inc., 1971.

Both teachers reportedly learned from this trade-off. One remarked that she had not realized a certain group of students never participated in her class until she sat at the back and watched them being taught by someone else.

The interns also reported benefits from observing. Under normal circumstances, interns rarely have the opportunity to see anyone teach a class other than their master teacher. As one intern put it, "By seeing other interns you get to see yourself with regard to your peer group—it is reassuring to know that you are not the only one making mistakes."

All participants liked the exposure to other methods of instruction and teaching styles. Teachers rarely have a chance to observe one another teaching—particularly if they are in self-contained classrooms. But even the teachers in the open-space school said that under usual conditions, they were too busy to observe their teammate adequately. Collegial evaluation gave them the chance not only to observe but to focus their observation using specific criteria.

The quality of feedback exchanged in the conferences was largely dependent on the quality of observations. The best observers were those guided by a few specific criteria that were appropriate to the particular activity they observed. They learned more from their observations and were better able to offer their partner concrete and useful information.

Conferences

Conferences require the ability to give constructive criticism without damaging egos or destroying long-term relationships. As our collegial evaluation program specifies, teachers in the pilot test exchanged feedback on three occasions: after each of the observation periods and at the wrap-up conference. In addition, they rated their strengths or weaknesses for each of the shared criteria on the self-evaluation form, which is similar to the observation form, making it easy to compare the two evaluations. In every case, participants were harder on themselves than their colleagues were.

The interns were much more willing than the elementary teachers to give low ratings to their colleagues and to give critical feedback on the observation form. Interns, by definition, are "people learning the skills of teaching," while certificated teachers (theoretically at least) already possess these skills. From this perspective, it is not surprising that interns were more comfortable offering written criticism than the elementary school teachers. During the conferences, however, teachers exchanged criticism and did more than pat one another on the back. Although they were reluctant to write down their negative comments, they were usually quite candid in their conferences.

An important purpose of the conferences is to develop specific strategies for improvement. Since the elementary school teachers worked together in the same classroom area, many of them identified problems that could be worked on cooperatively. For example, one pair agreed that the noise level in their area was occasionally too high and they discussed how, as members of a team, they could create a quieter learning atmosphere. Because these teachers worked together, they were motivated to help each other—to give feedback that would improve not only their individual teaching performance but the overall atmosphere of their classroom.

One teacher pointed out that a major difference between criticism during collegial evaluation and evaluations by an administrator was "the way criticism was phrased." We were continually impressed by the tact and diplomacy exhibited in the conferences. Criticisms were frequently presented as suggestions for alternative techniques. In one teacher's words, "Instead of having someone say, 'you should do this', a colleague was more likely to say, 'something that worked well for me was this technique.'" This approach not only was less threatening but was perceived as more legitimate. If the technique worked for a colleague, it was worth a try.

The interns' conferences emphasized diagnosis rather than specific recommendations. They spent more time and effort analyzing teaching strengths and weaknesses than the elementary school teachers did. Perhaps because of their relative inexperience, they did not have as many concrete suggestions to offer one another and instead devoted some time at each conference to brainstorming alternative teaching strategies.

Collegial evaluation provided positive reinforcement as well as constructive criticism. Suggestions for improvement were balanced with praise for effective teaching. Praise seemed to fill a very great need. As one teacher said, "When your colleague praises you, it means so much." Praise improves teaching by reinforcing successful practices, thus encouraging their frequent use. In school, teachers rarely receive praise from their colleagues because they are not observed or evaluated by them. Though the value of positive reinforcement in motivating pupils is universally recognized, this practice has seldom been extended to teachers—in spite of the fact that the importance of teachers' job satisfaction and faculty morale has long been recognized by teachers and administrators alike.

The feedback given in the conferences encompassed virtually every aspect of classroom activity. Teachers learned not only about their own performance but about the overall climate of their classroom. For example, one intern noted, "There was a warm, cooperative atmosphere in this classroom. It was created by allowing student work groups to sit together on pillows on the floor and emphasizing the importance of group evaluation for the task." Another intern summarized his feeling for a class by telling his partner, "People are noisy; that doesn't bother me. They are talking, getting excited, and having fun." On a more critical note, an art intern told his partner that clean-up period was "utter chaos" and suggested that students be assigned responsibilities for cleaning up after themselves.

Teachers also reported learning more about the behavior of particular students. One observer said of a self-directed project, "The autonomous kids go directly to work, but those who need a lot of teacher direction and support are left out." During a classroom discussion session, another observer noted, "While most students seem to be involved, a few appear to be untouched by the discussion." And during a lecture presentation another observer said, "A couple of students did not understand; they needed extensive clarification." These comments became catalysts for discussion in the conference. The observed teacher wanted to know which students were not autonomous, which were untouched by the discussion, and which needed further clarification. The partners then discussed ways to overcome these problems.

Some of the observations focused on problems of classroom discipline. Classroom control was more frequently discussed in conferences by interns than by teachers. Throughout the evaluation process, interns helped one another identify which students were creating problems and what might be done to improve classroom order. For example, one intern learned that "a small group of boys in the back are goofing off." Following the conference this small group was broken up and dispersed throughout the classroom.

After specific discipline problems had been openly discussed in the conferences, both interns and teachers often took steps to solve them. Overlooking a particularly noisy student is difficult when a colleague has identified the problem through systematic evaluation and provided a justification for action. For example, many interns reported a reluctance to openly chastise their students. They feared that any display of authority would squash independence or creativity, or perhaps more important that it would jeopardize their students' affection for them. But when a colleague says that a certain student is testing the limits of tolerance (and what's more, that the same student creates a similar problem in his or her own classroom), a teacher feels more justified in trying to find sound teaching techniques to bring that student into line.

Understandably, much of the feedback exchanged during conferences focused on the teacher's behavior in the classroom. Some discussions were directed at subject-matter presentation. Teachers gave each other useful information about the quality of materials used in lessons, the appropriateness of the language used in classroom presentations, the clarity of objectives and direction, and specific techniques for making their lessons more interesting. These comments ranged from general observations, such as "The material is going over the kids' heads," to more specific one, such as "Your explanation of chromatic half steps was a little complicated." Similarly, the suggestions for improvement ranged from general ones concerning the teacher's overall performance, such as "You should take at least a half hour to present material you are now covering in ten minutes," to very specific ones, such as "Why not give each student a copy of the keyboard to follow along during your explanation of chromatic half steps?"

The conferences also provided a forum for discussing teacher-student interaction, which was a matter of great concern to the participants, judging by both the criteria they chose for observing and the feedback they gave during conferences. A common observation was that a certain student or group of students was ignored. Many teachers wanted feedback concerning whether they used eye contact with everyone in their room, whether they called on different pupils rather than continually selecting the same ones, and whether they gave equal attention to students. One teacher learned that though she was successful in finding occasions to talk with all of her students individually about their art projects, most of her remarks were negative. In the conference her partner suggested that "students should get more reinforcement on the positive aspects of their work." Teachers continually praised one another for using positive reinforcement.¹ As one said, "You gave lots of 'warm fuzzies' this morning and it meant a lot to the kids."

On a more procedural note, participants found that holding conferences no more than two or three days after observations improved the quality of feedback. Similarly, the observation form (where ratings and comments on the colleague's performance are written) was more useful if it was completed immediately after observing. But most important, teachers reported that the quality of their conferences ultimately depended on the willingness of the partners to be reasonably honest with one another.

¹Teachers rarely told one another to be more critical of their students' work or to develop higher expectations for their students, either individually or as a class. They seemed to believe that each student should receive a lot of teacher warmth and approval regardless of his academic performance. We believe that this approach has serious flaws. Other research shows that students develop greatly inflated opinions of their academic skills in classrooms characterized by strong and uncritical teacher approval. Overstressing warmth and praise may have negative consequences, since it can lead students to have totally unwarranted beliefs about their academic skills. G.C. Massey, M.V. Scott, and S.M. Dornbusch, Racism without Racists: Institutional Racism in Urban Schools, Occasional Paper No. 8 (Stanford, Ca: Stanford Center for Research and Development in Teaching, 1975), pp.7-10. Reprinted from The Black Scholar, 7, No.3 (November 1975), pp. 10-19.

Self-Assessment and Student Questionnaire

Following the structure of our collegial evaluation program, several of those who participated in the pilot test distributed the student questionnaire to their classes and completed the self-assessment form as part of the evaluation process. The teacher questionnaire contains items parallel to the student questionnaire. These allow teachers to identify similarities and differences in their perceptions of themselves and their students' perceptions. For example, the teacher responds to the question, "How often do you encourage students to ask questions when they don't understand what's going on?" Students answer the similar question, "When you don't understand what's going on in this class, how often are you encouraged to ask questions?" Like the teacher, students use a five-point scale which ranges (for this question) from "always" to "never." After combining the student responses and computing a classroom average, the teacher can discover the level of agreement between his self-assessment and his students' assessment. Moreover, by looking at the distribution of responses, a teacher might find that some students "never" feel encouraged to ask questions, even though most students "usually" do. Both the classroom average and the distribution thus provide interesting and useful kinds of information.

The contribution of these questionnaires to the evaluation process was summarized by one teacher:

I believe that the student questionnaire was extremely valuable in providing information that I myself or a third person could not possibly provide adequately or accurately. The specific kinds of questions deal with those problems that cannot be readily observed. They focus on those students' personal and academic needs that are basic to learning.

One of the most striking results of the pilot test was the high level of agreement between teachers and students as shown by responses on their questionnaires. This similarity was not anticipated by the teachers. One teacher remarked, "I was very surprised to find that my own perceptions agreed fourteen out of twenty-one times (over 66%) with the average of the students. I think this proved that even though my class may not be the greatest one in the world, my students and I certainly agree on what it is." Another teacher said, "The questionnaires indicate that I

have a realistic understanding of my students' feelings toward the class and myself as a teacher."

Despite the general agreement, there were several items on the questionnaire that produced substantial disagreement between teachers and their students. These findings raised new questions and prompted teachers to investigate the underlying reasons for the discrepancy. For example, one teacher was surprised to find that on the average her students felt classwork was "usually" too fast and difficult. Her first interpretation was that she had overestimated her students' abilities. After looking more closely at the distribution of responses, she saw that almost as many students felt the work was "just right" as felt the work was "much too difficult." The second interpretation focused on the diversity of student ability in the classroom. To improve her teaching, she began to individualize instruction so that all of her students would be able to do some things well.

General disagreement was produced between the intern teachers and their high school students by another interesting question: "How important to you is having the teacher like you?" Secondary students rarely reported that this was either "extremely" or "very" important. The secondary interns seemed a little hurt and surprised by their students' indifference. This finding generated a very fruitful discussion among interns. It led to admissions that they were probably upset by this student report because they wanted so much to be liked by their own students. They had just assumed that liking was reciprocal. They confided to one another that wanting to be liked sometimes interfered with their better judgment as teachers. This conclusion was incorporated into their overall plans for improvement.

By comparison, elementary teachers were a little overwhelmed at their students' rating of their teacher's importance in their lives. Almost all elementary students said it was "extremely important" to be liked by their teacher. Of course, these veteran teachers had suspected that their students wanted their affection, but they had not known how strong or how widespread this feeling was. Such unanimity in their students' responses made them sensitive to a number of related behaviors in the classroom.

For example, after reviewing the questionnaire but prior to observation, one teacher noted about another, "Those kids are always touching you, and you never fail to respond."

In addition to insights gained from students' responses on each item of the questionnaire, teachers discovered that examining the responses on several items at once sometimes revealed interesting patterns. For example, one teacher discovered that her students reported being more confused than she had suspected. They agreed that the teacher's directions were unclear and that they were seldom encouraged to ask questions. She felt that their confusion might be alleviated if she took measures to clarify her directions and encouraged them to ask questions whenever they were confused.

Although anonymity was ensured on the student questionnaire, teachers and interns spent a lot of time guessing which students had given certain responses. The elementary teachers, who knew their students much better than the interns, seemed confident of their ability to make these guesses. When one student responded that he "never received good grades" even when he did "good work," the teacher said, "I know who that is, and he's right. We've got to start giving him some rewards for his efforts." The teacher was confident that this was the same student who responded that the teacher never let him know when he was doing "good work."

The participants agreed that maintaining anonymity was important if they wanted honest responses from students, but one lamented that "it would be valuable to know a particular student whose answers were radically different. It may be that this student is having difficult problems that I have overlooked or that are not obvious to me, and I would want to give him the special help that might be needed."

In the pilot test, one of the interns did a fine job of developing his own student questionnaire. He wanted to obtain specific information about his skills as a choir director. He learned that his conducting was "fairly easy to follow," but almost half of his students felt that he "stayed on one piece of music too long." Most of the choir liked the music "O.K.," with just a few liking it "a lot" or "not much." Only two students thought he looked like a "madman" when conducting. These items

provided an excellent supplement to the more general student questionnaire.

Student questionnaires provide teachers with information they cannot obtain elsewhere. Only students can tell a teacher whether or not they are interested and comfortable in the classroom. The problems students perceived were translated into specific criteria for the teacher's colleague to observe and were discussed in the conferences. The student assessment was a very valuable input that the teachers took into account in assessing their strengths and weaknesses and making plans for improvement.

Self-Assessment on Selected Criteria

In addition to the teacher questionnaire, participants completed a self-assessment form based on the criteria they had selected jointly with their partners. After their teaching was observed, this self-assessment could be compared with the observation form to help focus the conference on areas for improvement. Overall, participants were usually much more critical of themselves, both in ratings and in negative comments, than their colleagues were. They generally agreed with their partners' observations on areas of weakness, and most spent their conference in swapping ideas for improvement rather than in resolving disagreements.

A colleague's agreement was helpful in legitimatizing a teacher's perception of her strengths and weaknesses. For example, one teacher commented, "In discussion, I tend to rely on the same students who always have the answers, and I do not phrase open-ended questions to include everyone." When her colleague noted that "two boys spoke often, a few girls spoke occasionally, but no one else entered the discussion," her self-assessment was confirmed. A good part of their first conference focused on how she might increase student participation. In the second observation her colleague noted that "the discussion included more students and some who had not previously participated. You praised the newcomers-Good."

In her self-assessment another teacher noted a need for "some improvement" in lectures because she "relied too heavily on note cards." During the first observation her colleague identified the same area: "The organization and sequence of the lesson is good, but you occasionally stopped to

refer to notes." At the second observation the problem was not as severe and the colleague observed, "You relied on notes much less."

Of course, not all of the problems were so easily remedied. In a self-assessment one teacher reported the need "to project my voice." Her colleague noted, "Teacher's quiet voice tends to trail off" on the first observation form. In the second observation period the colleague reported, "Teacher's voice does not carry above sound of the slide projector." This is clearly a problem that needs to be addressed in that teacher's improvement plan.

The Improvement Plan

Developing a plan for improvement is the most important step of the collegial evaluation process. But the quality of each teacher's plan depends on how well the other steps have been carried out. The plan for improvement is formulated in a final "wrap-up" conference between the two partners. Each teacher integrates all the information he or she has received from self-assessment, student questionnaires, and peer evaluation, and presents his partner with a composite list of strengths and weaknesses. Together the teachers decide on the specific strategies each will use to improve their teaching performance in areas of weakness. In addition, they determine how they will evaluate the results of these strategies. Finally, they identify any resources they will need to carry out their improvement plan.

In our pilot test of collegial evaluation, the improvement plans spanned the whole range of teaching activities: presentation of subject matter, classroom control, motivation, student interest and involvement, positive reinforcement, and classroom organization and atmosphere. The improvement plans were based on evaluations that showed a remarkable amount of agreement between the teachers themselves, their colleagues, and their students. In most cases a teaching weakness identified by one of these sources was corroborated by the others.

For example, one teacher listed as an evaluation criterion, "Do not ignore any segment of the class concerning questions or needs--give attention equally." On the student questionnaire several students reported that

they were "seldom" or "never" encouraged to ask questions in class. On the basis of classroom observation the teacher's partner noted: "The less capable students are not involved, especially those at the back." As part of the plan for improvement, the teacher specified, "With the help of my peer, I will first identify those students whom I have ignored. I will make a point of talking to each of them every day. I'll keep a check list to make sure I spend some time with each of these children." Another teacher developed a plan to deal with a similar interaction problem in a different way. To encourage the nonparticipators at the back, she decided to rearrange the class and move the pupils at the back into the first two rows. She also said that she would "give those individuals who have not been participating responsibility for explaining things to the class and helping others with their work."

An intern chose as an evaluation criterion, "I present subject matter at a level appropriate to student ability." He was perplexed when most of his students reported on the questionnaire that they were confused by his explanations. Then his peer commented, "You use a lot of terms which go way over some of these kids' heads." In his improvement plan this intern listed a number of specific strategies to overcome the problem. Among these were: "I will try to define clearly all new terms which I use in class and be more careful to write these terms and their definitions on the board. I'll use pretests to determine pupil knowledge in the subject area. For those who do well on these tests I will design self-directed projects. This will leave me free to spend more time with the slow-achievers."

Some of the improvement plans called for relatively minor changes; others envisioned a major reorganization of the classroom and substantial changes in teacher behavior. Two of the elementary school teachers felt that they both needed to maintain a quieter learning environment. Such a concern is not atypical in open-space classrooms. After observing one another, they discovered that the noisiest time of the day came when they grouped their students by ability in math and language arts. The noise came from the "low ability" youngsters, and it prevented them and others from concentrating. As part of their improvement plan, the teachers

decided that the next year they would experiment with more heterogeneous groups.

Many of the identified weaknesses were not so difficult to remedy. For example, one art teacher, concerned about giving appropriate positive reinforcement for good work, benefited from his colleague's observation that he did not have any student work displayed in the classroom. He planned to "reserve a large space in the art room, school library, and hall display cases for the exhibition of student work." Another intern, whose problem was that he never had time to finish his lesson, decided to save a few minutes each period by letting students distribute and collect classroom materials rather than doing it himself.

For each of the specific strategies, teachers were asked to determine how they would assess their progress. Plans for assessment were as varied as improvement strategies. Teachers planning to improve their presentation of subject matter often relied on student cognitive outcomes as a measure of their success. The teacher mentioned above, who planned to explain and define new terms more carefully, listed as one indicator of progress the number of times students used the new terms in their essays.

Several teachers decided to use the student questionnaire as a post-test device to assess their improvement. Comparing the student response before and after the improvement plan was put into effect would help them assess their progress in such areas as motivating students, evaluating them, presenting material clearly, individualizing subject matter, displaying interest in students, and developing material appropriate to the students' level.

Almost all of the teachers planned to use collegial observation and conferences as a method of assessing their improvement. Many had already set up times to begin another round of observations with their colleagues. Others decided to change partners. The specific strategies for improvement would suggest new criteria for the next round of observations. One of the most gratifying results of the pilot test was that many of the participants considered our collegial evaluation program so useful that they planned to extend it throughout the school year. As one teacher said,

I need to have this kind of collegial evaluation on a regular basis. If my colleague evaluated me

throughout the year, she would have an understanding of the trends in my teaching and in a particular class and the evaluation would be even more helpful. She would be able to detect subtle problem areas that I may not be aware of. I could do the same for her and also continue to learn a lot by observing another teacher at work.

Conclusions

We began this discussion by criticizing traditional approaches to teacher evaluation and advocating collegial evaluation as an alternative. We summarized research revealing that teacher evaluation programs are all weak in one or more steps of the evaluation process. According to teachers and administrators we have interviewed, criteria for observation are usually vague or unknown, observations are infrequent, useful feedback is rare, and plans for teacher improvement are almost nonexistent. The experiences of teachers in the pilot test of our collegial evaluation program gave us some evidence for assessing this approach and comparing it with more traditional methods of evaluating teachers.

Most important, we learned that teachers can and will help each other perform better on their jobs. We also learned that teachers will take students' assessments of their teaching seriously and use them in developing plans for improvement.

We found that the most difficult step of our program was selecting criteria to serve as a basis for evaluation. But most teachers did select some criteria that were specific, observable, and meaningful to them. We also learned that thinking about their criteria helped teachers assess not only where they might need to improve but what their goals as teachers were.

We emphasized that the steps of the evaluation program are interdependent and that a weakness in any one of them would diminish the program's usefulness. This was especially apparent in reviewing improvement plans. If the criteria were specific, observable, and meaningful, if the observer was attentive and carefully reported observations to his or her colleague, and if the feedback exchanged was complete and honest, then the improvement plan generated by the pair of teachers was a thoughtful and practical blueprint for professional growth. The message is clear; teachers cannot

participate in this program in a half-hearted manner. If they are to use it as a means for improving their teaching, they must commit themselves to doing a thorough and careful job at every step.

Does collegial evaluation work? We believe the answer is yes. Based on our pilot test we have concluded that collegial evaluation is a useful approach to teacher evaluation in schools. On the whole, teachers reacted favorably to collegial evaluation, adapted the program to fit their unique circumstances, and gained new ideas for improving their teaching.

±

References

Dornbusch, S. M., and Scott, W. R. Evaluation and the Exercise of Authority. San Francisco: Jossey-Bass, 1975.

Marram, G. D., Dornbusch, S. M., and Scott, W. R. The Impact of Teaming and the Visibility of Teaching on the Professionalism of Elementary School Teachers. (Stanford Center for Research and Development in Teaching, Technical Report No. 33) Stanford University, December 1972.

Thompson, J. E., Dornbusch, S. M., and Scott, W. R. Failures of Communication in the Evaluation of Teachers by Principals. (Stanford Center for Research and Development in Teaching, Technical Report No. 43) Stanford University, April 1975.